
**IMPLEMENTATION OF MACHINE LEARNING
ALGORITHMS FOR HEART DISEASE PREDICTION**

Project report submitted in partial fulfillment of the requirement
for the degree of Bachelor of Technology

In

ELECTRONICS AND COMMUNICATION ENGINEERING

By:

Ishan Goyal (171035)
Tarun Arora (171034)

under the supervision

of

Dr. Harsh Sohal

To



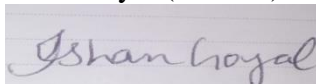
Department of Electronics and Communication Engineering
**Jaypee University of Information Technology, Wagnaghat, Solan,
Himachal Pradesh-173234**

Certificate
Candidate's Declaration

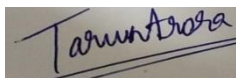
I as a consequence pronounce that the work added on this report named "execution of machine learning .Gaining knowledge of calculations for Heart Diseases Predection" in incomplete delight of the stipulations for the respect of the level of bachelor of era in electronics and communication submitted inside the department of electronics and communication engineering, Jaypee University of Information Technology, Waknaghat is my very personal real report paintings did over a length from jan 2021 to may 2021 below the oversight of Dr. Harsh Sohal.

The matter epitomized inside the file has no longer been submitted for the honor of some other degree or reputation.

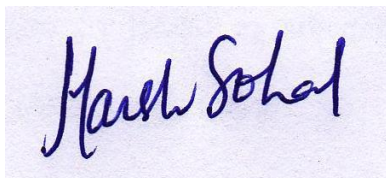
Ishan Goyal (171035)



Tarun Arora (171034)



This is to certify that the above statement made by the candidate is true to the best of my knowledge.



Dr. Harsh Sohal

Electronics and Communication

Dated: 19.5.2021

ACKNOWLEDGMENT

Working for the “Heart Diseases Analysis” project was interesting. We got to learn about Bayesian Networks, SVM, Logistic Regression and Decision Trees in Machine Learning, and to compare the accuracies and outcomes of each of these.

Special thanks to our supervisor Dr. Harsh Sohal for his guidance and advice on this project.

Also, we are very grateful to our college, lectures, and friends where they gave us enough time to complete this report and we would like to thank all other people who supported us in our project.

Thank you.

TABLE OF CONTENTS

Title.....	(i)
Certificate.....	(ii)
Acknowledgement.....	(iii)
List of figures.....	(v)
List of Tables.....	(v)
Abstract.....	(vi)
Chapter 1 Introduction.....	1
Chapter 2 Literature Survey.....	11
Chapter 3 Performance Analysis.....	28
Chapter 4 Conclusion.....	44
References.....	46

LIST OF FIGURES

Figure 1 Heart.....	3
Figure 2 Sample decision tree.....	9
Figure 4 Types of Machine Learning algorithms	14
Figure 5 Graphs depicting Dropout	16
Figure 6 flowchart of the three algorithms	18
Figure 7 Python libraries used in the project.....	25
Figure 8 Attributes in the dataset.....	23
Figure 9 DT root node.....	32
Figure 10 Analysis of Diabetes using DT	33
Figure 11 Training accuracy of KNN	34
Figure 12 Logistic regression plot	38

List of Tables

Table 1 Results.....	44
----------------------	----

ABSTRACT

"Coronary sickness Prediction" structure dependent upon keen appearance predicts the pollution of the client subject to the appearances that client gives as a pledge to the framework. The framework looks at the signs given by the client as data and gives the likelihood of the pollution as a yield.

Disease Prediction is done by executing the Naïve Bayes Classifier, Decision tree; Logistic Regression Naïve Bayes Classifier processes the probability of the ailment. Along these lines, typical conjecture exactness probability 80% is gotten.

Chapter 1

Introduction

1.1 Introduction

Perhaps the focal organs of the human body are a coronary heart. Probably the maximum unmistakable cardiovascular contaminations in india are the coronary bafflement. The heart siphons blood through the circulatory methodology of the frame. In all frame element the blood, oxygen is spouted with the aid of the circulatory methodology of the body and if the coronary heart doesn't fill in precisely actual to form, the whole human blood configuration will be collapsed. Consequently, enduring the coronary heart does not fill in precisely actual to form, it will impel a proper medical problem, it is able to even instigate annihilation.

1.2 Types of the Heart Disease:

The improvement of plaque can cripple the vessel surely through the course of time. The symptoms of the coronary heart assault:

1. Chest pain: the most placing signal of a respiration frustration is chest torture. It average happens assist the blockage of the coronary vessel of the body as a result of the plaque.

2. Arms torture: the torture consistently starts in the chest and circulate towards the arm essentially left arm.

Three. Sluggishness: this legitimization fatigues proposes direct errands emerge as significantly greater energetically to do.

Four. Over the pinnacle sweating: some other massive result is sweating.

Five. Bradycardia: in this, the affected person could have a good extra gradual heartbeat of 60 bpm.

6. Hypertension: on this the patient's pulse commonly going from a hundred-two hundred bpm.

7. A few different purposes behind the event of a coronary difficulty are way of life inclinations like smoking and sure dietary fashions. A typical thought that cannot placed forth an attempt no longer to be that extra than 17. Five million passings happen thinking about cardiovascular hassle within the whole global. In india, there may be greater than 30 million coronary soreness patients as of now. In india, more than 2 lake open heart physical games are completed every yr. The patients impacted by means of the breathing bafflement is filling in india is 20% to 30 % continuously. The development of the sensor community within the human checking gadget is more fitting from overdue years. In this enterprise, we show the use of sensor amassed information which typically made with the aid of the sensor how it'll in well known be carried out in gadget mastering reviews. Our errand is based at the information of the sensor which assembles the human heartbeat. By way of the use of the sensor facts and making use of it in machine getting to know appraisal we can assume the coronary sickness.

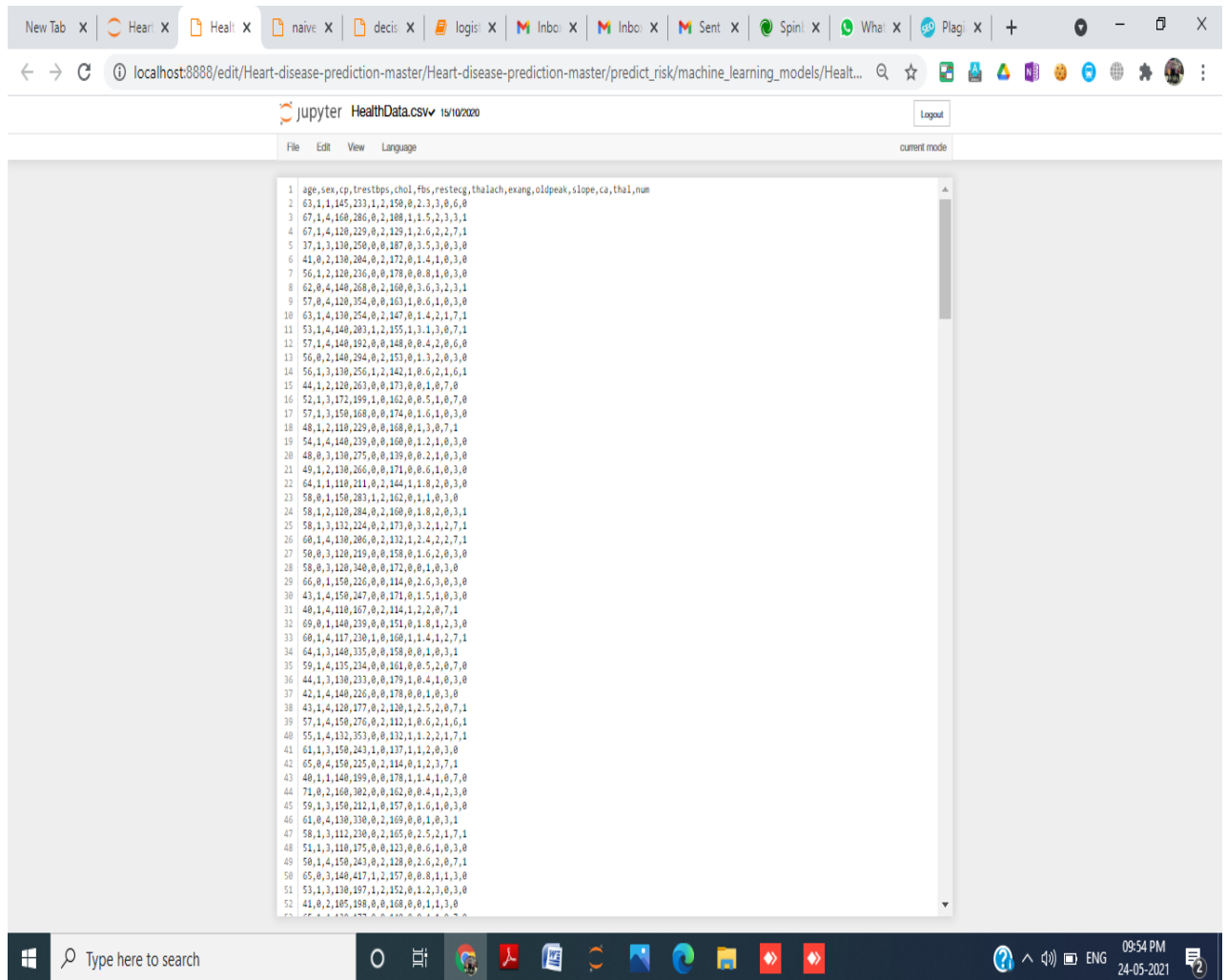
Proposed system

In this mission, we suggest a design which can be utilized for each coronary infection seeing and assessment. In this project, the proposed structure shakes the blueprint of the beat and the suspicion for coronary sickness may be seen via anticipated statistics. Thinking about the expansion or diminishing the hrv the version will anticipate the chances of the disease to occur.

The Dataset

```
#! (utf-8)
import pandas.core.frame.DataFrame
RangeIndex: 302 entries, 0 to 301
Data columns (total 18 columns):
 age                302 non-null float64
 sex                302 non-null float64
 chest_pain         302 non-null float64
 blood_pressure     302 non-null float64
 serum_cholesterol  302 non-null float64
 resting_blood_sugar 302 non-null float64
 electrocardiogram  302 non-null float64
 max_heart_rate     302 non-null float64
 noncardiol_angina  302 non-null float64
 st_segmentation   302 non-null float64
 kcp               302 non-null float64
 no_of_vessels      299 non-null float64
 thal              302 non-null float64
 diagnosis          302 non-null int64
 dtypes: float64(13), int64(1)
memory usage: 33.2 KB
```

Figure: 'missing values' of the dataset.



Symptoms:

1. Chest torment, chest snugness, chest urgent issue and chest uneasiness (angina)
2. Windedness.
3. Affliction, deadness, shortcoming or frigidity to your legs or arms if the blood Vessel in the ones portions of your body is restrained.
4. Real annoyance, jaw, throat, upper mid-location or returned

Treatment:

The objectives of treatment are settling the condition, controlling indications over the long haul, and giving a fix whenever the situation allows. Stress decrease, diet, and way of life changes are key in overseeing coronary illness, yet the backbones of traditional consideration are medications and medical procedure.

1.2 Problem statement

The general goal of my paintings could be to expect precisely with few assessments and features the presence of coronary infection. Characteristics considered structure the critical cause for assessments and give genuine effects. Loads extra information credits can be taken but we can likely expect with no longer many ascribes and quicker effectiveness, the danger of having coronary infection made depending on specialists' intuition and revel in instead of on the facts wealthy statistics hidde. This training prompts unwanted inclinations, mistakes and intense clinical costs which impact the character of administration gave to sufferers.

1.3 Objective

We centre on considering each characteristic in our dataset and applying the previously mentioned calculations for the forecast of the result. Thusly, we would have the option to examine which characteristic generally assumes a huge part in the assurance of coronary illness. Additionally, what least worth of that quality would bring about a positive result?

Additionally, this undertaking would assist us with recognizing that utilizing which calculation we can anticipate the result with greatest exactness and accuracy. By this, we can tell the advantages and disadvantages of every calculation utilized.

We will complete a thorough and similar examination by breaking down the after-effects of every calculation to figure out which calculation can be the best fit for an underlying expectation of the event of Diseases in a patient.

Algorithms used for the analysis of diabetes are:

Naïve Bayes Classifier:

This order procedure uses probability theory for portrayal of data. This methodology works dependent on Bayes' theory. The essential information on Bayes' speculation expresses: the probability of an event might be adjusted as new data is introduced.

This classifier is called gullible due to its assumption that every quality of data under idea is self-sufficient of each other.

This is a gathering of ML plots that use measurements self-governance. This procedure is reasonably easy to form as wells as execute more capable than other Bayes' strategies.

Support Vector Machines:

This plan alludes to a coordinated ML strategy that might be used for both relapse and characterization calculations. However, this is to a great extent applied to characterization calculations. For SVM, our point is to orchestrate each information as a spot with facilitates in a n-dimensional area in which n alludes to the quantity of qualities in our information where every property has a place with a particular organize. After this, our goal is to carry out gathering through looking a hyperplane what isolates the two classes outstandingly fine.

Backing vectors allude to data which is close the hyperplane and effect the situation just as a prologue to the hyperplane. Using the vectors, we extend the edge inside the classifier. Deleting the help vectors will change the spot of the hyperplane.

Decision Trees: Decision tree are used methodology designed for deriving plus order through depiction using center points and internodes. The primary and inward centers comprise of the trials which are used to detach the events with different features.

The inward hub itself is the outcome of a quality experiment. Leaf centers connote the case changeable. Exhibits an illustration of a Decision tree.

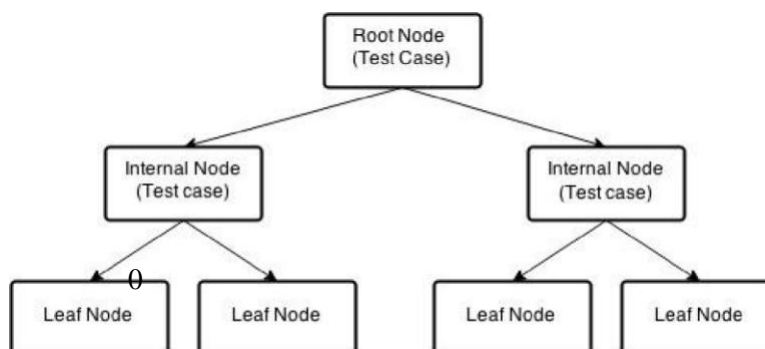


Figure 2 Sample decision tree

1.4 Logistic Regression

Strategic relapse are real Machine learning calculation masterminds a data with allowing for outcome factor happening exceptional terminations or endeavors make a logarithmic row that remembers them.

An articulation —Logistic are taking from Logic work i.e. used technique of gathering otherwise supposed grouping. Did like utilize calculation exploration since gives an answer for the grouping issue which the fundamental point of this task for example I should arrange whether the patient is diabetic dependent on the given boundaries. The outcomes got by this strategy are as a Yes or a No.

1.5 Organization

Chapter 1, we had talked about Heart disease illness or difficulties looked with individuals. Likewise, brought up plans that we would use to examine disease undertaking Report.

In Chapter 2, we had talk about exploration Papers we had alluded improve comprehension of our undertaking. The papers essentially center around methods utilized in AI and different explores completed out.

In Chapter 3, we had refer to potential necessities that are the equipment and programming framework utilize and where are we carry out it alongside the libraries needed alongside insights regarding the stage utilized.

In Chapter 4, we would talk about exhaustively the calculations that anticipate the result and viability of our outcome. Executions and the after-effects of the yields had been examined.

In Chapter 5, we had give ends have gotten investigation extent of this undertaking.

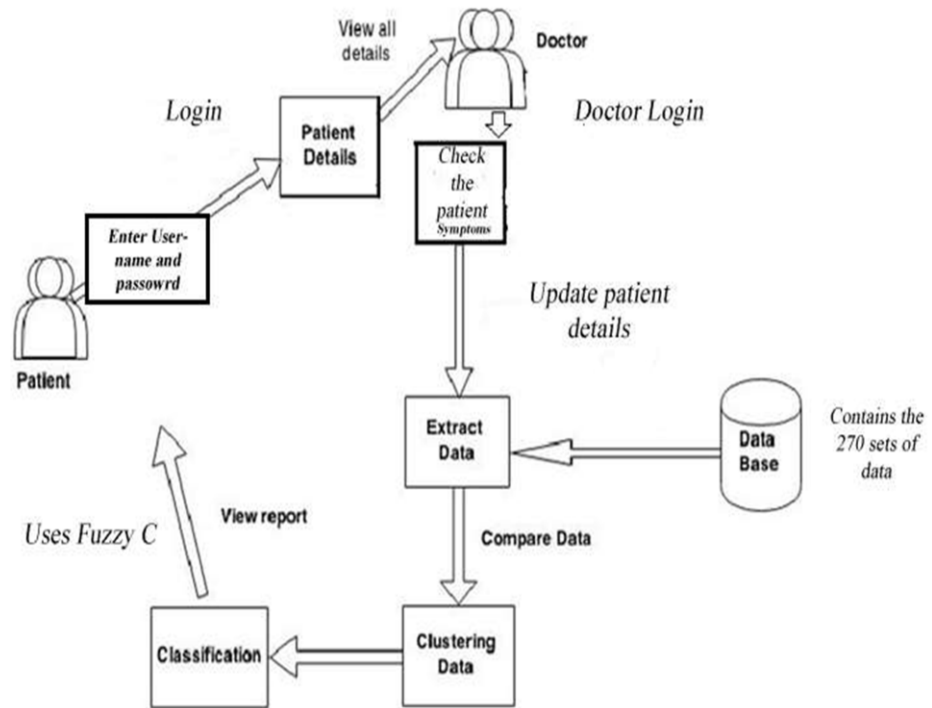


Fig.1.1 Architecture Of The System

Chapter 2

Literature Survey

A) Systematic Literature Survey

2.1 "AI applications in disease visualization and forecast"

Generally, researchers utilized the methodology of screening at an underlying stage even before any of the manifestations supposedly identified a sort of threatening development. Recently, the existences of various presentations inside the flood of medication yet most proficient results, likewise expectations are delivered by ML calculations. These techniques can discover and recognize models and associations inside complex datasets while they satisfactorily expect future outcomes having a place with the infection type. Also, the benefits and disservices of each AI technique joined have been outlined.

The crucial device utilized is ANN. ANNs manage a meeting of depiction or the difficulty of enduring plan. Those are prepared for making the yield in view of mixing among figures. The resulting one is the DT. A decision tree is an essential and big philosophy used in portrayal. It carries a plan of a tree where a center infers facts; yield statistics is tended to by way of the leaf.

Backing Vectors partition the information into two gatherings dependent on the individual information highlights through a hyper plane. The last one is BN. Bayesian Network follows a methodology of discovering likelihood rather than direct assessment.

Thinking about the assessment of the result, the dataset they utilized alongside the utilization of different frameworks or ways to deal with picking ascribes can bring up apparent instruments for the harmful development space.

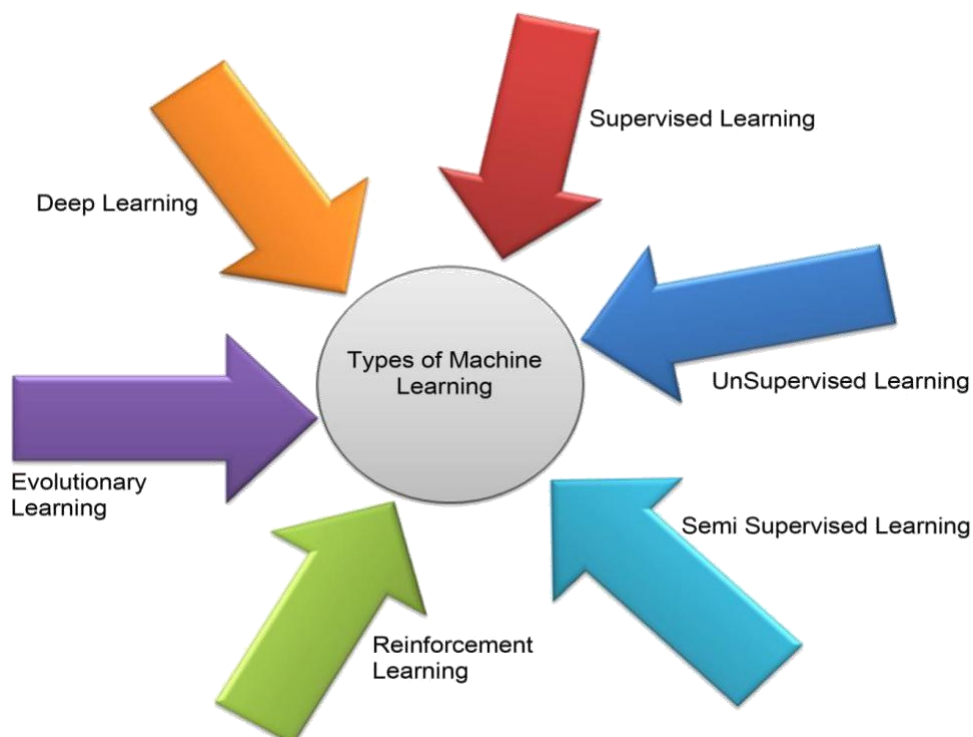
2.2 Survey of Machine Learning Algorithms for Disease Diagnostic

Creators: Meherwar Fatima

Maruf Pasha

These techniques have gotten fundamental for mechanized investigation. Subsequent to using straightforward condition organs may fundamentally not be shown absolutely. In this way, plan affirmation from an overall perspective incorporates acquiring from these, models. Inside the flood of medication, plan affirmation and AI ensure the upgraded exactness for the investigation of affliction.

Figure: Types of Machine Learning algorithms



MLL computations for illness examination. Experts have recognized that ML computations work outstandingly in the examination of different ailments. This paper represents the accompanying sicknesses broke down by AI: Liver, heart, diabetes, hepatitis, and dengue.

On account of heart illnesses, SVM gives the most exact outcomes that are of 94.6%. Also, support vectors utilize deciding the most suitable attributes and this methodology brings about an accuracy of 84.1% which is less contrasted with what is appeared by SVM. Guileless Bayes calculation and choice tree was utilized for the discovery of diabetes.

Discovered to be the most productive with an accomplishment of 94%. Notwithstanding better finding, this calculation calls attention to diminished paradoxes. One of the disadvantages of utilizing this methodology was that it required an immensely gigantic dataset. Various examiners utilized different blends of various ML strategies to recognize the liver sickness; however the most exact results was recorded by FT tree (Feature choice tree). This calculation set aside lesser effort for its advancement in contrast with the others with the precision of 95.1%.

Dengue disease is among the real irresistible afflictions. In its determination, the Rough set hypothesis demonstrated to have the most extreme accuracy. It is talented to regulate weakness, uproar just as lacking qualities. Assurance of credit connects with the classifier to beat substitute models. In the space of examination for hepatitiis, a neural organization explicitly FFNN showed greatest accuracy with the estimation of

This survey paper presents a few instances of ML calculations that have demonstrated to create the best results because of their ability of separating the

attributes exactly. In addition, it illustrates a pool of instruments which are appeared to have advanced in the field of computerized reasoning.

2.4 "Comparing the computational intricacy and accuracy of classification calculations"

They led an assessment to think about the intricacy of arrangement calculations. They contrasted two calculations an information mining and the Pioneer classifier calculation. Their paper delineates the previously mentioned way to deal with One of the essential tasks of information mining is a gathering of data.

The 3 algos talked about go for describing which choice tree is ideal just as simple to understand and interpret. The SL in Quest method can manage mathematical just as subjective qualities. In this calculation, a GINI Index is made where each mathematical worth has n ascribes. In the last methodology, for example the one improving the effectiveness of SLIQ was grown for the most part to oust the disadvantages of SLIQ. This plan is a quicker methodology as the minimum number of calculations are completed

2.5 —An observational investigation of the innocent Bayes classifier"

Creators: Irina Rish

The Naive Bayes classifier essentially unravels learning by tolerating that the features are self-sufficient given class. In spite of the fact that self-governance is normally a helpless speculation, all things being equal Naive Bayes consistently fights well with continuously complex classifiers.

The objective of this examination was to understand the data credits who impact the implimentation of gullible Bayes. Monte Carlo reenactments have

been used in this way to deal with permit a conscious examination of course of action precision for a couple of classes of heedlessly made issues.

They have taken apart the impact of the assignment entropy in the gathering batch; showing up low energy incorporates apportionments yields incredible accuracy. Unquestionably, it has additionally been exhibited .

The examination reasons that in spite of its impossible independence assumption, the guileless Bayes order strategy is incredibly amazing eventually since its portrayal decision may often be correct paying little mind .

Amazingly, the accuracy of this classifier isn't clearly associated with the degree of feature conditions assessed as the class-unforeseen regular information between the features. Maybe, a prevalent pointer of precision is the deficiency of information that highlights. In any case, further trial and speculative mull over is needed to more probable grasp the association between those information hypothetical estimations and the lead of guileless Bayes.

Further headings moreover fuse the examination of gullible Bayes on the logical application that has almost deterministic conditions, portraying various locale of innocent Bayes optimality, besides, concentrating the effect of various information highlights on the Bayes batch.

2.6"Prediction Accuracy of various procedures

Creators: K.M. Al-Aidaros, A.A. Bakar and Z.

Othman has directed the exploration for the best clinical finding mining strategy. For this creator contrasted Naïve Bayes and five different classifiers for example Strategic For this, 15 certifiable clinical issues from the UCI AI store (Asuncion and Newman, 2007) were chosen for assessing the presentation, all things considered. In the trial it was discovered that NB beats different calculations in 8 out of 15 informational collections so it was inferred that the prescient precision brings about Naïve Baeyes is superior to different methods.

Table 1- Predictive Accuracy of Bayes and other Technique

Medical Problems	NB	LR	K*	DT	NN	ZeroR
Breast Cancer wise	97.3	92.98	95.72	94.57	95.57	65.52
Breast Cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Echocardiogram	95.77	94.59	89.38	96.41	93.64	67.86
Liver Disorders	54.89	68.72	66.82	65.84	68.73	57.98
Pima Diabetes	75.75	77.47	70.19	74.49	74.75	65.11
Haebberman	75.36	74.41	73.73	72.16	70.32	73.53
Heart-c	83.34	83.7	75.18	77.13	80.99	54.45
Heart-statlog	84.85	84.04	73.89	75.59	81.78	55.56
Heart-b	83.95	84.23	77.83	80.22	80.07	63.95
Hepatitis	83.81	83.89	80.17	79.22	80.78	79.38
Lung Cancer	53.25	47.25	41.67	40.83	44.08	40
Lymphpgraphy	84.97	78.45	83.18	78.21	81.81	54.76
Postooperative Patient	68.11	61.11	61.67	69.78	58.54	71.11
Primary tumor	49.71	41.62	38.02	41.39	40.38	24.78
Wins	8\15	5\15	0\15	2\15	1\15	1\15

(Al-Aidaros, Bakar, & Othman, 2012)

Chapter 3

Execution Analysis

3.1 All Round evaluation

On this part we will look at the entirety of the calculations which have been applied in this assessment undertaking. Close by the important definitions and the thoughts concerning the exams, the execution of the nearly identical has been appeared. Additionally, it will damage down computations applied for setting apart diabetes:

3.1.1 Naïve Bayes Algorithm:

To choose associations among the signs, end, and prescriptions of diabetes, distinctive mining or yet they go with a couple of shortcomings like bluntness and dynamically immense time for learning the outcomes. BN deals with these burdens in its own specific habits by taking out a growing number of repeated assumptions. Furthermore, this arrangement may moreover be utilized for a giant enlightening assortment continuously.

This procedure is completed using the going with condition:

Execution of Naïve Bayes:

```
In [4]: print("{0:0.2f}% data is in training set".format((len(x_train)/len(pdata.index)) * 100))
print("{0:0.2f}% data is in test set".format((len(x_test)/len(pdata.index)) * 100))
```

```
69.92% data is in training set
30.08% data is in test set
```

```
In [5]: from sklearn.naive_bayes import GaussianNB # Gaussian algorithm from Naive Bayes

# create model
diab_model = GaussianNB()

diab_model.fit(x_train, y_train.ravel())
```

```
Out[5]: GaussianNB(priors=None)
```

```
#Gaussian Naive Bayes
model = train_model(X_train, y_train, X_test, y_test, GaussianNB)
```

```
Train accuracy: 85.38%
```

```
Test accuracy: 86.81%
```

3.1.5 Decision **tree** algorithm:

This procedure gives a solid way to deal with the assessment and arrangement of Diabetes illness. Each hub of such a tree is given by registering the biggest information increment inside every one of the properties, yet when a specific property creates a dubious last result, its driving branches are stopped, and the last worth is distributed to that unit.

The underneath figure shows part of the execution of the Decision tree

```
In [7]: from sklearn import tree

In [8]: d_tree = tree.DecisionTreeClassifier(criterion='gini',
      splitter='best', random_state = 0)
      d_tree.fit(x_train, y_train)

Out[8]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
      max_features=None, max_leaf_nodes=None,
      min_impurity_decrease=0.0, min_impurity_split=None,
      min_samples_leaf=1, min_samples_split=2,
      min_weight_fraction_leaf=0.0, presort=False, random_state=0,
      splitter='best')

In [9]: import pydotplus

In [11]: dot_data = tree.export_graphviz(d_tree, out_file=None,
      feature_names=features, class_names=class_label)
      #Draw graph
      graph = pydotplus.graph_from_dot_data(dot_data)
      graph.write_pdf('diab-tree.pdf')
```



```
In [12]: y_pred = d_tree.predict(x_test)
```

```
In [14]: ## Accuracy of the decision tree model
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_pred)
score = accuracy_score(y_test, y_pred)
print(cm)
print(score)
```

```
[[123  34]
 [ 30  44]]
0.7229437229437229
```

Execution of KNN:

```
In [7]: from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(diabetes.loc[:,
                                                    diabetes.columns != 'Outcome'],
                                                    diabetes['Outcome'],
                                                    stratify=diabetes['Outcome'],
                                                    random_state=66)
```

```
In [8]: from sklearn.neighbors import KNeighborsClassifier
training_accuracy = []
test_accuracy = []

neighbors_settings = range(1, 11)
for n_neighbors in neighbors_settings:

    knn = KNeighborsClassifier(n_neighbors=n_neighbors)
    knn.fit(X_train, y_train)

    training_accuracy.append(knn.score(X_train, y_train))

    test_accuracy.append(knn.score(X_test, y_test))

plt.plot(neighbors_settings, training_accuracy, label="training accuracy")
plt.plot(neighbors_settings, test_accuracy, label="test accuracy")
plt.ylabel("Accuracy")
plt.xlabel("n_neighbors")
plt.legend()
plt.savefig('knn_compare_model')
```

```
In [8]: knn = KNeighborsClassifier(n_neighbors=9)
knn.fit(X_train, y_train)

print('Accuracy of K-NN classifier on training set: {:.2f}'.format(knn.score(X_train, y_train)))
print('Accuracy of K-NN classifier on test set: {:.2f}'.format(knn.score(X_test, y_test)))
```

Accuracy of K-NN classifier on training set: 0.79
Accuracy of K-NN classifier on test set: 0.78

for further queries continue expanding the quantity of closest neighbors up ideal score 80% and 90% on the preparation set and the test set separately.

3.1.4 Logistic Regression

Calculated is an order calculation it can be utilized when a straight out reaction variable is required. It very well may exist utilized decide connection qualities likelihood of a particular yield. The articulation work that is used in this system for gathering or supposed arrangement. I like to utilize this calculation in my exploration since it gives an answer for the grouping issue which the primary point of this venture for example I should arrange whether the patient is diabetic dependent on the given boundaries. The outcomes got by this strategy are as a Yes or a No.

An explanation of the strategic capacity will give us the explanation of Logistic Regression.

Strategic capacity takes esteems somewhere in the range of 0 to 1 motivation behind Sigmoid Function.

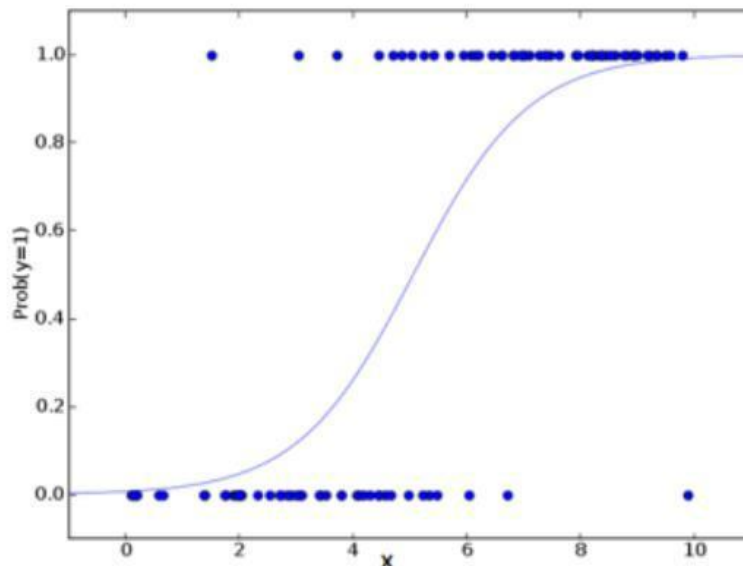


Figure: Logistic regression graph

Let's consider t as linear function in a univariate regression model.

$$t = \beta_0 + \beta_1 x$$

So the Logistic Equation will become

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Now, when logistic regression model come across an outlier, it will take care of it.

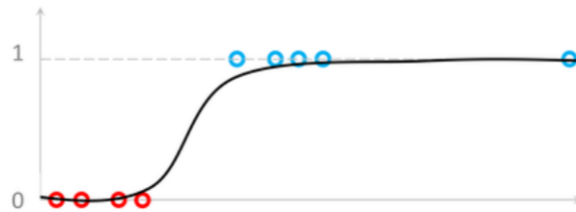


Figure 12 logistic regression model

Implement of LG Regression:

```
# Logistic Regression
model = train_model(X_train, y_train, X_test, y_test, LogisticRegression)

Train accuracy: 85.85%
Test accuracy: 85.71%
```

Right while the critical backslide estimation is applied without the regularization restriction c , we get the readiness set accuracy as 85.Eighty five% and the test set rating as 85.71%.

```
In [11]: logreg001 = LogisticRegression(C=0.01).fit(X_train, y_train)
print("Training set accuracy: {:.3f}".format(logreg001.score(X_train, y_train)))
print("Test set accuracy: {:.3f}".format(logreg001.score(X_test, y_test)))
```

```
Training set accuracy: 0.700
Test set accuracy: 0.703
```

```
In [12]: logreg100 = LogisticRegression(C=100).fit(X_train, y_train)
print("Training set accuracy: {:.3f}".format(logreg100.score(X_train, y_train)))
print("Test set accuracy: {:.3f}".format(logreg100.score(X_test, y_test)))
```

Training set accuracy: 0.785
Test set accuracy: 0.766

We should envision the coefficients learned by the models with the three distinct settings of the regularization parameter C . More grounded regularization ($C= 0.001$) pushes coefficients increasingly more toward 0. Assessing the plot more intently, we can likewise observe that highlight "DiabetesPedigreeFunction", for $C=100$, $C=1$ and $C=0.001$, the coefficient is certain. This shows high "DiabetesPedigreeFunction" include is identified with an example being "diabetes", notwithstanding which model we take a gander at.

3.1.5 Support Vector Machine (SVM)

A support vector system (svm) is a discriminative classifier formally depicted by means of a keeping apart hyperplane. Alongside these traces, given a directed and named dataset for getting geared up (managed studying), the method yields a super hyperplane which puts collectively new perspectives. In 2D area, this hyperplane is a line proscribing a aircraft in quantities in which every elegance lays in either side. This method can once more be used for both – amassing and backslide, anyway generally for request.

The motives why this calculation has been taken into consideration for this area are: svm features admirably with the unmistakable fringe of section, is profoundly powerful in in which no. Of measurements is larger than data passages, and it

makes use of a detachment of making plans facilities desire potential, sooner or later it's miles further memorable capable. Implement of Support Vector Machine:

```
In [22]: #svm
```

```
In [23]: from sklearn.svm import SVC
svc = SVC()
svc.fit(X_train, y_train)
print("Accuracy on training set: {:.2f}".format(svc.score(X_train, y_train)))
print("Accuracy on test set: {:.2f}".format(svc.score(X_test, y_test)))
```

```
Accuracy on training set: 1.00
Accuracy on test set: 0.65
```

Here we observe that we obtain a perfect training set score whereas the test set gives only 65% accurate results.

From this, we can incur that our features are not aligned on a similar scale. Tuning parameters esteem for AI calculations viably improves the model execution. So we re-scale all these features:

```
In [24]: from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.fit_transform(X_test)
svc = SVC()
svc.fit(X_train_scaled, y_train)
print("Accuracy on training set: {:.2f}".format(svc.score(X_train_scaled, y_train)))
print("Accuracy on test set: {:.2f}".format(svc.score(X_test_scaled, y_test)))
```

```
Accuracy on training set: 0.77
Accuracy on test set: 0.77
```

Since SVM streamlining happens by limiting the decision vector w , the ideal hyperplane is affected by the size of the information highlights and it's along these lines suggested that information be institutionalized preceding SVM model preparing.

Scaling the data parameters had an immense effect! Presently we are really underfitting, where preparing and test set execution are very comparable yet less near 100% precision. So now we will probably increment the value of gamma or C to set into a more intricate model.

Greater the estimated value of gamma will attempt to correct fit the according to pre the data set for example speculation mistake and cause an over-fitting issue.

```
In [25]: svc = SVC(C=1000)
          svc.fit(X_train_scaled, y_train)
          print("Accuracy on training set: {:.3f}".format(
              svc.score(X_train_scaled, y_train)))
          print("Accuracy on test set: {:.3f}".format(svc.score(X_test_scaled, y_test)))
```

Accuracy on training set: 0.790

Accuracy on test set: 0.797

When we set the value of C=100, the model improves and ends up in giving out 79.7% test set precision.

4.2 Results

The table beneath records every one of the recorded upsides of the yields that were gotten all through the examination. The principal segment contains the name of the calculation. The factors section facts the elements that were taken into consideration in that calculation to shift the consequences. The closing two sections list the instruction and take a look at set exactness rankings in my view.

```

y_Class_pred=classifier.predict(X_test)

#checking the accuracy for predicted results

from sklearn.metrics import accuracy_score
accuracy_score(Y_test,y_Class_pred)
    
```

Out[15]: 0.7763157894736842

```

In [16]: # Making the Confusion Matrix

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, y_Class_pred)
    
```

```

In [17]: #Interpretation:

from sklearn.metrics import classification_report
print(classification_report(Y_test, y_Class_pred))
    
```

	precision	recall	f1-score	support
0	0.76	0.89	0.82	44
1	0.88	0.62	0.78	32
micro avg	0.78	0.78	0.78	76
macro avg	0.78	0.76	0.76	76
weighted avg	0.78	0.78	0.77	76

```

In [18]: #ROC

from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(Y_test, classifier.predict(X_test))
fpr, tpr, thresholds = roc_curve(Y_test, classifier.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc='lower right')
plt.savefig('log_roc')
plt.show()
    
```



Downloads | Heart-disease-prediction-master | naive_bayes.py | decision_tree.py | logistic_jupyter

localhost:8888/edit/Heart-disease-prediction-master/Heart-disease-prediction-master/predict_risk/machine_learning_models/naive...

Jupyter naive_bayes.py 15/02/2021 Logout

File Edit View Language Python

```
1 # Importing the libraries
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import pandas as pd
5
6 # Importing the dataset
7 dataset = pd.read_csv('HealthData.csv')
8 X = dataset.iloc[:, :-1].values
9 y = dataset.iloc[:, 13].values
10
11 # Handling missing data
12
13 from sklearn.preprocessing import Imputer
14 imputer=Imputer(missing_values='NaN', strategy='mean', axis=0)
15 imputer=imputer.fit(X[:, 11:13])
16 X[:, 11:13]=imputer.transform(X[:, 11:13])
17
18
19
20 # Splitting the dataset into the Training set and Test set
21 from sklearn.model_selection import train_test_split
22 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.15, random_state = 0)
23
24 # Feature Scaling
25 from sklearn.preprocessing import StandardScaler
26 sc = StandardScaler()
27 X_train = sc.fit_transform(X_train)
28 X_test = sc.transform(X_test)
29
30 # EXPLORING THE DATASET
31 import seaborn as sn
32 sn.countplot(x='num', data=dataset)
33 dataset.num.value_counts()
34
35
36 # Fitting Naive Bayes to the Training set
37 from sklearn.naive_bayes import GaussianNB
38 classifier = GaussianNB()
39 classifier.fit(X_train, y_train)
40
41 from sklearn.externals import joblib
42
43 #filename = 'naive_bayes_model.pkl'
44 #joblib.dump(classifier, filename)
45
46 # Predicting the Test set results
47 y_pred = classifier.predict(X_test)
48
49 # ACCURACY SCORE
50 from sklearn.metrics import accuracy_score
51 accuracy_score(y_test, y_pred)
52
53 # Making the Confusion Matrix
```

Type here to search | 07:33 PM | 24-05-2021

Downloads | Heart-disease-prediction-master | decision_tree.py | logistic_jupyter

localhost:8888/edit/Heart-disease-prediction-master/Heart-disease-prediction-master/predict_risk/machine_learning_models/decis...

jupyter decision_tree.py 15/10/2020

```

21 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
22
23 # Feature Scaling
24 from sklearn.preprocessing import StandardScaler
25 sc = StandardScaler()
26 X_train = sc.fit_transform(X_train)
27 X_test = sc.transform(X_test)
28
29 #EXPLORING THE DATASET
30 import seaborn as sns
31 sns.countplot(x='num', data=dataset)
32 dataset.num.value_counts()
33
34 # Fitting Decision Tree Classification to the Training set
35 from sklearn.tree import DecisionTreeClassifier
36 classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
37 classifier.fit(X_train, y_train)
38
39 from sklearn.externals import joblib
40 filename = 'decision_tree_model.pkl'
41 joblib.dump(classifier, filename)
42
43 # Predicting the Test set results
44 y_pred = classifier.predict(X_test)
45
46 #ACCURACY SCORE
47 from sklearn.metrics import accuracy_score
48 accuracy_score(y_test, y_pred)
49
50 #CONFUSION MATRIX
51 from sklearn.metrics import classification_report, confusion_matrix
52 cm=confusion_matrix(y_test, y_pred)
53
54 #Interpretation:
55 from sklearn.metrics import classification_report
56 print(classification_report(y_test, y_pred))
57
58 #ROC
59 from sklearn.metrics import roc_auc_score
60 from sklearn.metrics import roc_curve
61 logit_roc_auc = roc_auc_score(y_test, classifier.predict(X_test))
62 fpr, tpr, thresholds = roc_curve(y_test, classifier.predict_proba(X_test)[:,1])
63 plt.figure()
64 plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
65 plt.plot([0, 1], [0, 1], 'r--')
66 plt.xlim([0.0, 1.0])
67 plt.ylim([0.0, 1.05])
68 plt.xlabel('False Positive Rate')
69 plt.ylabel('True Positive Rate')
70 plt.title('Receiver operating characteristic')
71 plt.legend(loc='lower right')
72 plt.savefig('log_roc')
73 plt.show()

```

Type here to search | 07:33 PM 24-05-2021

Downloads | Heart-disease-prediction-master | logistic_jupyter

localhost:8888/notebooks/Heart-disease-prediction-master/Heart-disease-prediction-master/predict_risk/notebooks/logistic_jupy...

jupyter logistic_jupyter Last Checkpoint Last Friday at 11:38 (AutoSaved)

```

In [6]: Exploring the dataset
import matplotlib.pyplot as plt
def plot_histograme(dataset, feature, row, col):
    plt.figure(figsize=(8, 8))
    for i, feature in enumerate(features):
        ax=plt.subplot(row,col,i+1)
        dataFromFeature=dataset[feature].hist(bins=20,color='darkblue')
        ax.set_title(feature+' Distribution',color='darkred')
    plt.tight_layout()
    plt.show()
plot_histograme(dataset,dataset.columns[6:7])

```

```

In [7]: dataset.num.value_counts()
import seaborn as sns
sns.countplot(x='num', data=dataset)

```

Type here to search | 07:28 PM 24-05-2021

Chapter 4

Conclusion

4.1 Conclusion

A table beneath records every one of the recorded upsides of the yields that were acquired all through the examination. The Factor segment records the elements that were calculation shift outcomes. The three sections list the training Our study uncovers the adequacy of every calculation utilized. Most likely not one but rather numerous calculations have end up being comparable in giving out the for all intents and purposes indistinguishable exactnesses for our troublesome affirmation and the field of investigation. idn't wind up being much incredible as they give is the after-effects of just 86% and 79%(approx.). We saw that for some, the estimations, setting the right boundaries is huge for acceptable execution. On the off chance that we plot the element significance of every one of the pre-owned calculations.

6.2 Future Scope

Subjects were researched suggest assumption and evasion are as restored and strengthened by ML applications, while "security and disillusionment area" had generally explored. Through my eyes, investigators in this field should continue abusing the latest redesigns in ML and to go along with them with progression. Various assessments reported exact supposition and divulgence devices that confirmation to improve the pioneer's assets for present and future drugs.

Additionally, test set exactness scores separately.

At the point when everything the use of Artificial intelligence frameworks begin from data driven strategies that addition from colossal datasets. The ability to assemble information from singular diabetic patients has provoked a

move in diabetes the board systems; moreover, structures that need admittance to beneficial data will defy liberal snags.

An ethical perils related with the appearance of individual data should moreover be investigated. For example, the unquestionably visit usage of prosperity applications and the possible use of gadgets subject to AI by protection organizations could provoke isolation or the dismissal (or both) of specific inhabitants from prosperity organizations. Exploration in this field ought to continue hope to discover the odds and inclinations of applying artificial intelligence strategies in diabetes the board that different these approaches set up procedures.

References

1. Fatima, M. and Pasha, M., —Survey of Machine Learning Algorithms for Disease Diagnostic. | Journal of Intelligent Learning Systems and Applications, (2017), 9, 1-16.
2. K. Kourou et al./Computational and Structural Biotechnology Journal 13 (2015), 8–17
3. Acknowledgment for the dataset: Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265).
4. Do we Need Hundreds of Classifiers to Solve Real-World Classification Problems? Journal of Machine Learning Research 15 (2014) 3133-3181
5. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, —Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929-1958
6. I. Rish, —An empirical study of the naive Bayes classifier, T.J. Watson Research Centre (2001)
7. Lakshmana Ayaru, Petros-Pavlos Ypsilantis, Abigail Nanapragasam, Ryan Chang-Ho Choi, Anish Thillanathan, Lee Min-Ho, Giovanni Montana. Prediction of Outcome in Acute Lower Gastrointestinal Bleeding Using Gradient Boosting
8. Clare Martin, Antonio Martinez-Millana, Andrew Stranieri, Klerisson Paixao, Maurice Mulvenna, and Francisco Nuñez-Benjumea. —Clare Martin, Antonio Martinez-Millana, Andrew Stranieri, Klerisson Paixao, Maurice Mulvenna, and Francisco Nuñez-Benjumea (2018)
9. [The Naive Bayes Algorithm in Python with Scikit-Learnstackabuse.com](#)
10. [What is K-Nearest Neighbor \(K-NN\)? - Definition from Techopedia](#)www.techopedia.com
11. [What is Machine Learning? A definition - Expert System](#)www.expertsystem.com
12. <https://www.researchgate.net/publication/271850951>
13. [Introduction to Decision Trees \(Titanic dataset\) | Kaggle](#)www.kaggle.com
14. How to explain gradient boosting <https://explained.ai/gradient-boosting/>
15. Logistic Regression- <https://towardsdatascience.com/tagged/logistic-regression>
16. Support Vector Machine- <https://towardsdatascience.com/tagged/svm>